Nucleic Acids Research, 2001, Vol. 29, No. 1 **199–201**

The RESID Database of protein structure modifications and the NRL-3D Sequence–Structure Database

John S. Garavelli*, Zhenglin Hou, Nagarajan Pattabiraman[1] and Robert M. Stephens[1]

National Biomedical Research Foundation, Protein Information Resource, Washington, DC 20007, USA and [1]Advanced Biomedical Computing Center, SAIC, National Cancer Institute–Frederick, Frederick, MD 21702, USA

Received October 2, 2000; Revised and Accepted October 27, 2000

ABSTRACT

The RESID Database is a comprehensive collection of annotations and structures for protein post-translational modifications including N-terminal, C-terminal and peptide chain cross-link modifications. The RESID Database includes systematic and frequently observed alternate names, Chemical Abstracts Service registry numbers, atomic formulas and weights, enzyme activities, taxonomic range, keywords, literature citations with database cross-references, structural diagrams and molecular models. The NRL-3D Sequence–Structure Database is derived from the three-dimensional structure of proteins deposited with the Research Collaboratory for Structural Bioinformatics Protein Data Bank. The NRL-3D Database includes standardized and frequently observed alternate names, sources, keywords, literature citations, experimental conditions and searchable sequences from model coordinates. These databases are freely accessible through the National Cancer Institute–Frederick Advanced Biomedical Computing Center at these web sites: http://www.ncifcrf.gov/RESID, http://www.ncifcrf.gov/NRL-3D; or at these National Biomedical Research Foundation Protein Information Resource web sites: http://pir.georgetown.edu/pirwww/dbinfo/resid.html, http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html

INTRODUCTION

The NRL-3D Sequence–Structure Database was originally developed and tested by Krish Namboodiri, Nagarajan Pattabiraman, Alfred Lowrey and Bruce P. Gaber in 1990 at the Laboratory for the Structure of Matter and the Center for Bio/Molecular Science and Engineering at the Naval Research Laboratory, Washington, DC and at the Department of Biochemistry and Molecular Biology, Georgetown University, Washington, DC (1). In 1992 the NRL-3D Database was adapted to the multi-database access programs of the National Biomedical Research Foundation Protein Information Resource (PIR) and produced in a format similar to the PIR-International Protein Sequence Database by David George and John S. Garavelli of the National Biomedical Research Foundation, Washington, DC, and Masami Kusunoki of the Institute for Protein Research, Osaka University, Osaka, Japan (2). The original objectives of the NRL-3D Database were to provide a means for performing sequence searches on the models represented in the ATOM records of the Protein Data Bank (PDB), and to provide a means for extracting atomic coordinate information for models matching segments of sequence. Although more annotation information that had been parsed from the PDB files and standardized was added to NRL-3D entries in 1995, the capability for extracting the model coordinates was lost. SEQRES sequence records were subsequently added to PDB files. However, those sequences did not necessarily correspond to the sequences modeled and they were represented in the ATOM records with varying residue index labels. The NRL-3D Database still provided the most reliable means for searching the sequences actually represented in the models of PDB files. In 1998 the NRL-3D Sequence Structure Database previously produced on a periodic basis by the PIR began to be produced at the facilities of the Advanced Biomedical Computing Center of the National Cancer Institute–Frederick, Frederick, MD.

In 1993, the RESID Database of Protein Structure Modifications began as a database of standardized features representing modified amino acids residues reported in the PIR-International Protein Sequence Database (3). The RESID Database was first distributed on CD-ROM accompanying the PIR in 1995, and in 1998 it was made available on the web first with graphical and later with model components. In the latter part of 2000 production of the RESID Database of protein structure modifications also began at the Advanced Biomedical Computing Center of the National Cancer Institute–Frederick. The RESID Database was designed: (i) to document covalent binding sites, modified sites and cross-links with bibliographic, structural descriptions, keywords and other annotations; (ii) to furnish more detailed chemical information than is possible in a protein sequence database, and to enable users to recognize when different authors are using synonymous descriptions of previously described features; (iii) to provide an adaptable mechanism for calculating the molecular weights of modified proteins and their peptide fragments; and (iv) to be accessible through the Internet and database access programs with search capabilities and display of chemical structures.

RESID is the only publicly available, distributed database comprehensively documenting more than 260 structural and regulatory modifications as well as active site prosthetic modifications, and providing both visual display and molecular models for these protein post-translational modifications.

*To whom correspondence should be addressed. Tel: +1 202 687 2121; Fax: +1 202 687 1662; Email: jsgaravelli@earthlink.net

DATABASE DESCRIPTIONS

The RESID Database includes entries for the 22 alpha-amino acids known to be genetically encoded, including N-formyl methionine and selenocysteine, for the three ambiguous 'residues' represented in the IUPAC standard single-letter code, and for more than 260 other residues either predicted or observed in proteins arising through natural, post-translational modification of encoded amino acids. A few molecular structures that are known not to exist but have appeared in the literature are included for comprehensiveness and quality assurance. The developmental XML distribution format provides for the inclusion of both artificially produced modifications that are commonly encountered in mass spectrographic analysis and for amino acids that occur naturally in non-genetically encoded peptides.

Information in RESID database entries includes: dates for database entry and modification of text and structure; a systematic chemical name and Chemical Abstracts Service registry number for the free residue; frequently observed alternate names; the atomic formula and weight; original amino acids with difference formulas and weights; enzyme activities producing the modification; indicators for N-terminal, C-terminal or peptide chain cross-link modifications; literature citations, keywords and feature table representations for the modification in the PIR-International and SWISS-PROT protein sequence databases. Future releases will include indicators for whether the modification is artificial or natural with the observed taxonomic range. The RESID Database maintains concurrent cross-references to the PIR-International Protein Sequence Database, the Chemical Abstracts (CAS), the MEDLINE citation database and PDB. (CAS Registry Numbers are copyrighted by the American Chemical Society and used with permission of the Chemical Abstracts Service of the American Chemical Society.) Structural diagrams are presented in GIF format, and molecular models in PDB format are provided for use with widely available Web display programs such as RasMol (4) and Chime (MDL Information Systems, Inc., San Leandro, CA).

Web users are provided capabilities for searching lists of modifications arising from particular encoded amino acids, for searching for modifications based on either the molecular weight of the modified residue or the difference in the molecular weight produced by the modification. These weights are calculated using either chemical-average or monoisotopic molecular in order to facilitate their identification by mass-spectroscopy (5,6).

The NRL-3D Sequence–Structure Database is derived from the three-dimensional structure of proteins deposited with the Research Collaboratory for Structural Bioinformatics Protein Data Bank (7). The NRL-3D Database includes standardized and frequently observed alternate names, the title name in the PDB entry, biological source, keywords, literature citations, experimental conditions and searchable sequences from model coordinates. Separate chains in PDB entries are presented as separate entries in NRL-3D. However, peptide fragments are assumed to occur whenever sequence adjacent α-carbon atoms are separated by more than 4.8 Å, and these are now represented by the IUPAC standard sequence punctuation '/' separating the peptide fragments rather than by placing the fragments in separate entries as in earlier versions of NRL-3D. The developmental XML distribution format provides for the inclusion of sequences from other structure depositories such as the BioMagResBank, Repository for Data from NMR Spectroscopy on Proteins, Peptides and Nucleic Acids (8), and for the comparative presentation of sequences in the SEQRES records, as well as from protein sequence databases.

AVAILABILITY AND ACCESS

The RESID and NRL-3D databases are frequently updated and distributed in their original NBRF format, and in developmental XML format. Database entries may be accessed through the National Cancer Institute–Frederick Advanced Biomedical Computing Center at these web sites: http://www.ncifcrf.gov/RESID, http://www.ncifcrf.gov /NRL-3D; or at these National Biomedical Research Foundation Protein Information Resource web sites: http://pir.georgetown.edu/pirwww/dbinfo/pir-resid.html, http://pir.georgetown.edu/pirwww/dbinfo/pir-nrl3d.html. The RESID Database may be searched by entry code or other unique identifier, by name, citation, keyword or feature text search, by molecular weight search, or from selection lists based on encoded amino acids. The NRL-3D Database may be searched by entry code or other unique identifier, by name, citation, keyword or feature text search, or by FASTA, BLAST or peptide pattern matching.

The RESID Database is copyrighted. Both the NRL-3D and RESID databases are distributed free with no license required. It is appreciated if authors cite this article or the introductory announcements of the RESID (9) or NRL-3D (1) databases.

SUBMISSIONS AND REVISIONS

John S. Garavelli invites the submission of information for new entries or for the revision of existing entries in the RESID Database. Those wishing to submit material may do so by email directed to the author's attention at jsgaravelli@earthlink.net. New database entries are assigned unique access codes, which may be cited in publications.

ACKNOWLEDGEMENTS

J.S.G. acknowledges Mr David J. Miller for his help as research assistant. Enhancement of the RESID database was supported by NSF grant DBI-9808414. Hosting and support for the RESID and NRL-3D databases is kindly provided by the Advanced Biomedical Computing Center, SAIC, National Cancer Institute–Frederick, Frederick, MD. RESID and NRL-3D are registered marks of J.S.G. PIR is a registered mark of the National Biomedical Research Foundation.

REFERENCES

1. Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) NRL_3D: a sequence–structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Seq. Data Anal.*, **3**, 387–405.
2. Garavelli,J.S. (1993) The NRL_3D Database: A Tool for Sequence Conformation Study. *Protein Sci.*, **2** (Suppl. 1), 126.
3. Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.-W., Orcutt,B.C., Pfeiffer,F., Tsugita,A., Vinayaka,C.R., Xiao,C., Yeh,L.-S.L. and Wu,C.(2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32.

4. Sayle,R.A. and Milner-White,E.J. (1995) RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
5. Yates,J.R.,III, Eng,J.K., McCormack,A.L. and Schieltz,D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, **67**, 1426–1436.
6. Kelleher,N.L. (2000) From primary structure to function: biological insights from large-molecule mass spectra. *Chem. Biol.*, **7**, R37–R45.
7. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 214–218.
8. Seavey,B.R., Farr,E.A., Westler,W.M. and Markley,J.L. (1991) A Relational Database for Sequence-Specific Protein NMR Data. *J. Biomol. NMR*, **1**, 217–236.
9. Garavelli,J.S. (1993) A Database of Protein Structure Modifications. *Protein Sci.*, **2** (Suppl. 1), 133.